MACHINE INTELLIGENCE GARAGE

CATAPULT Digital

Ethics Framework

Machine Intelligence Garage Ethics Committee has created an Ethics Framework consisting of seven concepts, along with corresponding questions intended to inform how they may be applied in practice.

The Ethics Framework is a highly practical tool for individuals and organisations developing AI-enabled products and services who want to build value-aligned technologies with positive effects whilst avoiding negative consequences.

The Ethics Framework was created by Machine Intelligence Garage's Ethics Committee and launched in 2018. This emphasis is on questions rather than high-level principles because questions help illuminate where principles should be considered in practice, and questions do not assume a universal 'correct' answer.

In the research phase of this project, we found a multitude of useful references and used these to derive the seven concepts. As a result, the framework is closely aligned with the more recent ones developed by the High-Level Expert Group on Artificial Intelligence of the European Commission in 2019 and with the OECD's Principles on AI and the Beijing AI Principles.

We use the Ethics Framework in the consultations between the Ethics Advisory Group (a subset of the Ethics Committee) members and Machine Intelligence Garage startups, and it is updated regularly as a result of this feedback loop. The Ethics Framework is licensed under a Creative Commons Attribution 4.0 International Licence and we encourage feedback via appliedAIethics@digicatapult.org.uk

How to use this framework?

The framework consists of seven concepts with a corresponding list of questions for each. The questions are intended to illuminate the many contexts in which that ethical concept might be relevant to the business or project. Not all questions will be relevant at all times, and many questions will not have an immediate or clear answer.

Consult the framework around key milestones in the project. The right time to start thinking about the questions is at the earliest stages of company growth. Consider current, near and mid-future potential effects. The framework will help to characterise ethical opportunities and potential risks and to be open and clear both internally and externally about how these are evaluated and managed.

We advise companies to consult the following seven points:

Suggested action plans against these principles will be informed by product maturity and adoption. We will therefore suggest that companies consider current, near and mid-future potential effects.

1. Be clear about the benefits of the product or service

While it is important to consider the risks of new technologies, this should be done in the context of expected benefits. The benefits should be clear, likely, and outweigh potential, reasonable risks. They should be evaluated for different user groups and for any affected non-user groups (especially when there are competing values or interests between these groups), and with consideration of plausible future trends or changes (for example greater compute capacity or a solution coming to dominate the market).

- What are the goals, purposes and intended applications of the product or service?
- Who or what might benefit from the product/service?
- Consider all potential groups of beneficiaries, whether individual users, groups or society and environment as a whole.
- Are those benefits common to the application type, or specific to the technology or implementation choices?
- How to monitor and test products or services meet these goals, purposes and intended applications?
- How likely are the benefits and how significant?
- How to assess what the benefits are?
- How are these benefits obtained by the various stakeholders?
- Can the benefits of the product/service be demonstrated?
- Might these benefits change over time?
- What is the team's position on making (parts of) the products/services available on a non-commercial basis, or on sharing AI knowledge that would enable more people to develop useful AI applications?

2. Know and manage the risks

Safety and potential harm should be considered, both in consequence of the product's intended use, and other reasonably foreseeable uses. For example, the possibility of malicious attacks on the technology needs to be thought through. Attacks may be against safety, security, data integrity, or other aspects of the system, such as to achieve some particular decision outcome. As with benefits, assessment of risks can be in respect of individuals (not just users), communities, society and environment and should consider plausible future trends or changes.

- What might be the risks of other foreseeable uses of the technology, including accidental or malicious misuse of it?
- What are the potential groups at risk, whether individual users, groups or society and environment as a whole?
- Is there currently a process to classify and assess potential risks associated with the use of the product or service?
- Who or what might be at risk from the intended and non-intended applications of the product/service? Consider all potential groups at risk, whether individual users, groups, society as a whole or the environment.
- Are those risks common for application area or technology, or specific to the technology or implementation choices?
- How likely are the risks, and how significant?
- Is there a plan to mitigate and manage the risks?
- How to communicate the potential risks or perceived risks to users, potentially affected parties, purchasers or commissioners?
- How do third-parties or employees report potential vulnerabilities, risks or biases, and what processes are in place to handle these issues and reports?

2. Know and manage the risks

- How to tell if bias has been created or reinforced with the system?
- Are there any customers or use cases that would not be worked with as a result of assessing potential risks? How are these decisions made and documented?

3. Use data responsibly

Compliance with legislation (such as GDPR) is a good starting point for an ethical assessment of data and privacy. However, there are other considerations that arise from data-driven products, such as the aptness of data for use in situations that were not encountered in the training data, or whether data contains unfair biases, that must be taken into account when assessing the ethical implications of an AI product or service.

Data may come in many forms: as datasets, through APIs, through labour (such as microtasking). The value exchange between those who provide the data (or label it), directly or otherwise, and the company, should be considered for fairness. If data is used from public sources (for example open data collected by a public body or NGO) the company should consider whether it may contribute back or support the work of ongoing data maintenance, perhaps by providing cleaned or corrected data.

- How was the data obtained, was consent obtained (if required)?
- Is the data current?
- Is the training data appropriate for the intended use?
- Is the data pseudo-anonymised or de-identified? If not, why not?
- Is the data use proportionate to the problem being addressed?
- Is there sufficient data coverage for all intended use-cases?
- What are the qualities of the data (for example, is the data coming from a system prone to human error?)?
- Are potential biases in the data examined, well-understood and documented and is there a plan to mitigate against them?
- Is there a process for discovering and dealing with inconsistencies or errors in the data?
- What is the quality of the data analysis? How much uncertainty/error is there? What are the consequences that might arise from errors in analysis and how can these be mitigated?

3. Use data responsibly

- Can it clearly be communicated how data is being used and how decisions are being made?
- What systems are in place to ensure data security and integrity?
- Are there adequate methods in place for timely and auditable data deletion, once data is no longer needed?
- Can individuals remove themselves from the dataset? Can they also remove themselves from any resulting models?
- Is there a publically available privacy policy in place, and to what extent are individuals able to control the use of data about them, even when they are not users of the service or product?
- Are there adequate mechanisms for data curation in place to ensure external auditing and replicability of results, and, if a risk has manifested itself, attribution of responsibility?
- Can individuals access data about themselves?
- Is the data made available for research processes?

4. Be worthy of trust

For a technology or product to be trusted it needs to be understood, fit-for-purpose, reliable and competently delivered. Companies should be able to explain the purpose and limitations of their solutions so that users are not misled or confused. There should be processes in place to monitor and evaluate the integrity of the system over time, with clarity over what the quality measures are, and how they are chosen. Care must be taken to operate within the company's areas of competence, and to actively engage with third-party evaluation and questions. Things can go wrong, despite best efforts. Companies should put in place procedures to report, investigate, take responsibility for, and resolve issues. Help should be accessible and timely.

- Within the company, are there sufficient processes and tools built in to ensure meaningful transparency, auditability, reliability and suitability of the product output?
- Have the limitations of experience on the system being built been acknowledged and how can these reflect on the system in place? What steps are being taken to address these limitations?
- Is the nature of the product or technology communicated in a way that the intended users, third parties and the general public can access and understand?
- Are (potential) errors communicated and their impact explained?
- Does the company actively engage with its employees, purchasers/commissioners, suppliers, users and affected third parties so that ethical (including safety, privacy and security) concerns can be voiced, discussed, and addressed?
- Does the company work with researchers where appropriate to explore or question areas of the technology?
- Is there a process to review and assure the integrity of the AI system over time and take remedial action if it is not operating as intended?
- If human labour has been involved in data preparation (eg image labelling by Mechanical Turk workers) have the workers involved been fairly compensated?

4. Be worthy of trust

- Who is accountable if things go wrong? Are they the right people? Are they equipped with the skills and knowledge they need to take on this responsibility?
- What is/are the quality or standards to which the product/technology must conform (fro example academic; peer review, technical), what are the reasons for choosing the particular standards; and what does the company propose to do to maintain such standards?
- In order to engender trust, are there customers, suppliers or use cases that the company should choose not to work with? How are these decisions made and documented?
- Does the company have a clear and easy to use system for third party user or stakeholder concerns to be raised and handled?
- Is adequate training provided in the safe and secure use of the product or service to all of the operators, customers and/or users?
- Has there been consideration as to how to embed ethics within the organisation?
- Has there been any consideration as to how to embed integrity and fair dealing in the culture?
- How would a person raise a concern with the company?
- To inform the processes and culture, could approaches to mentors or consult innovation hubs be made?

5. Promote diversity, equality and inclusion

We will prioritise companies that can demonstrate that they value and actively seek diversity, equality and inclusion. Companies should consider the impact and utility of their product for individuals, larger groups and society as a whole, including its impact on widening or narrowing inequality, enabling or constraining discrimination, and other political, cultural and environmental factors. Diverse teams that are representative and inclusive are smarter, provide higher returns, and help create products and services that work for a greater number of people in society.

- Are there processes in place to establish whether the product or service might have a negative impact on the rights and liberties of individuals or groups? Please consider:
  – varied social backgrounds and education levels
  – different ages
  – different gender and/or sexual orientation
  – different nationalities or ethnicity
  – different political, religious and cultural backgrounds
  – physical or hidden disabilities.

- What actions can be taken if negative impacts are identified?

- Social impact can be difficult to demonstrate: Have the processes that can enable demonstration of the positive impact of the product or service been considered?

- Has putting in place a diversity and inclusiveness policy in relation to recruitment and retention of staff been considered?

- Has how to balance the specific responsibilities of a startup against other factors such as cost and freedom of choice for users been considered?

- Are potential biases in the data and processes examined, well-understood and documented and is there a plan to mitigate against them?

- Where do hiring practices and building culture fit in? For instance, are ethical questions raised at interviews? Are any principles/risk considerations communicated to new hires?

- Does the company have a diversity and inclusiveness policy in relation to recruitment and retention of staff?

6. Be open and understandable in communications

Companies must be able to communicate clearly the benefits and potential risks of their products and the actions they have taken to deliver benefits and avoid, minimise, or mitigate the risks. They must ensure that processes are in place to address the concerns and complaints of users and other parties, and that these are transparent. We believe that effective communication, when coupled with a principled approach to ethical considerations, is a competitive advantage, and will lead to progress even when hard moral issues are on the line. Conversely, poor communication, and a lack of attention to the social and ethical environment for doing business, can result in adverse public reactions, direct legal repercussions as well as mounting regulation, and hence increased costs and higher rates of failure.

- Does the company communicate clearly, honestly and directly about any potential risks of the product or service being provided?
- What does it communicate and when?
- Does the company communicate clearly, honestly and directly about the processes in place to avoid, minimise or mitigate potential risks?
- Does the company have a clear and easy to use system for third party/user or stakeholder concerns to be raised and handled?
- Are the company's policies relating to ethical principles available publicly and to employees? Are the processes to implement and update the policies open and transparent?
- Does the company disclose issues other than the product for example projects, studies and other activities funded by the company or which the company may work in conjunction, or otherwise be involved, with; the major sources of data and expertise that inform the insights of AI solutions and the methods used to train those systems and solutions?
- Has a communication strategy and process if something goes wrong been considered?

7. Consider the business model

Integrity and fair dealing should be an integral part of organisational culture. Companies should consider what structures and processes are being employed to drive revenue or other material value to the organisation as certain business models or pricing strategies can result in discrimination. Where possible and appropriate, companies should consider whether part of the product, service or data can be made available to the public.

- What kind of corporate structure best meets the company's needs? As well as the traditional company limited by shares, there are a variety of 'social enterprise' alternatives, including community interest company, co-operative, B-Corp and company limited by guarantee. Are any of these of interest?
- Data exchange: are free services in exchange for user data provided? Are there any ethical implications for this? Do users have a clear idea of how the data will be used, including any future inking/sale of the data?
- What happens if the company is acquired? For example, what happens to its data and software?
- Pricing: have differential prices been considered? Are there any ethical considerations regarding the pricing strategy? Are there any vulnerable groups to which lower prices may be offered?
- Data philanthropy: is there data that others could (e.g. charities, researchers) use for public purpose benefits?
- Is integrity and fair dealing embedded in the organisational culture?
- Has the environmental impact of development/deployment of the technology been considered?
- Is environmental impact considered when choosing suppliers? Have less energy-intensive options been considered?

References

In the process of drafting our ethics principles, we have researched and found useful the following publications:

- AI Code – the House of Lords Artificial Intelligence Committee "AI in the UK: ready, willing and able?" report recommends a cross-sector AI Code be established, which can be adopted nationally, and internationally. The code has five suggested principles.

- AI Hippocratic Oath – an article by Oren Etzioni, CEO of the Allen Institute for Artificial Intelligence. He edits the Hippocratic Oath sworn by generations of doctors to suggest an equivalent oath that AI practitioners can take to highlight their ethical commitments.

- Asilomar AI Principles – these 23 principles, developed at a conference held by the Future of Life Institute, have been signed by 1274 AI/robotics researchers and 2541 others (27 June 2018) including many household names in the world of AI and machine learning. Future of Life Institute says, "We hope that these principles will provide material for vigorous discussion and also aspirational goals for how the power of AI can be used to improve everyone's lives in coming years."

- Athena Swan – a charter that recognises and celebrates good practice towards the advancement of gender equality, established and managed by the British Equality Challenge Unit in 2005.

- Centre for Democracy and Technology – not-for-profit organisation that is championing online civil liberties and human rights, driving policy outcomes that keep the internet open, innovative, and free. CDT has created a tool to make one think about various challenges that could arise when designing, building, testing or implementing an algorithm, The tool; Blog piece about the tool.

- DataKind – is a not-for-profit organisation that brings together top data scientists with leading social change organisations to collaborate on cutting-edge analytics and advanced algorithms to maximize social impact. Their UK Principles establish what their community should abide by when working on data-for-good projects.

- Datasheets for Datasets – a paper written by authors from Microsoft Research, University of Maryland, Cornell University, Georgia Tech and AI Now Institute proposing to document datasets for greater transparency and accountability. They describe how datasheets for datasets will facilitate better communication between dataset creators and users, and encourage the machine learning community to investigate how a dataset was created, what information it contains, what tasks it should and shouldn't be used for, and whether it raises any ethical or legal concerns.

- DCMS – Data Ethics Framework (published Jun 2018) sets out seven principles for how data should be used in the public sector in order to "help maximise the value of data whilst also setting the highest standards for transparency and accountability when building or buying new data technology". The associated Data Ethics Workbook sets out the questions that should be considered against each of the principles.

- Doteveryone – (forthcoming) Responsible Technology Product Management Toolkit. "We are currently in the process of developing a number of assessment tools, which product teams can work through to help them examine and evaluate how they handle the 3Cs (context, consequences, and contribution) of responsible technology in real time during the development cycle. The form of the assessments ranges from checklists to step-by-step information mapping to team board games." Doteveryone is seeking help to road test the 3C model.

- EPSRC Principles of Robotics – five rules and seven principles for regulating robots in the real world. These "highlight the general principles of concern expressed by" a group of experts convened to "discuss robotics, its applications in the real world and the huge amount of promise it offers to benefit society" with the intention that they can "inform designers and users of robots in specific situations".

- Ethical OS Toolkit – released by Institute for the Future and Omidyar Network, the Ethical OS Toolkit is, "a toolkit designed to help technologists envision the potential risks and worst-case scenarios of how their technologies may be used in the future so they can anticipate issues and design and implement ethical solutions from the outset."

- The Future of Computing Academy (part of the Association for Computing Machinery) has proposed that the computer science community change its peer-review process to ensure that reviewers assess claims of impact as well as intellectual rigour. Hence researchers should think about and disclose any possible negative societal consequences of their work in their papers.

- Google's AI Principles – in June 2018, Google published seven principles to guide its work in AI research, product development and business decisions.

References

- Information Accountability Foundation – this global information policy think tank helps frame and advance data protection law and practice through accountability-based information governance. Providing tools for establishing legitimacy in big data projects. Noteworthy publications are:
  - Unified Ethical Frame for Big Data Analysis – theoretical basis for legitimacy
  - Big Data Assessment Framework and Worksheet – assessment framework for establishing legitimacy

- It Speaks – a research report produced by King's College London and ReFiG in Canada with the aim of providing solutions to the ethical problem of bias that exists in artificial intelligence language data sets.

- Open Data Institute – an independent, non-profit, non-partisan company focused on building an open, trustworthy data ecosystem. The ODI Data Ethics Canvas is a tool designed to help identify potential ethical issues associated with a data project or activity.

- Partnership on AI – a multistakeholder organisation that brings together academics, researchers, civil society organisations, companies building and utilising AI technology, and and other groups working to better understand AI's impacts. Partnership on AI has developed a set of Thematic Pillars that provide guidance on principles for developing AI.

- RAEng – Diversity and Inclusion Progression Framework. This tool helps engineering and science professionals (and soon startups) self-assess and improve their diversity and inclusion (D&I) maturity.

- Royal Society – the Society's fundamental purpose is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity.

- The Royal Society's Data Management and Use: Governance in the 21st Century provides a comprehensive review on the needs of a 21st century data governance system.

- Technology Strategy Board – the "Responsible Innovation Framework for commercialisation of research findings". This framework was developed for use assessing synthetic biology applications, but clearly has the potential to inform responsible technology more widely.

- The Universal Declaration of Human Rights – Drafted by representatives with different legal and cultural backgrounds from all regions of the world, the Declaration was proclaimed by the United Nations General Assembly in Paris on 10 December 1948 (General Assembly resolution 217 A) as a common standard of achievements for all peoples and all nations. It sets out, for the first time, fundamental human rights to be universally protected and it has been translated into over 500 languages.

MACHINE INTELLIGENCE GARAGE

CATAPULT Digital

Digital Catapult is the UK's leading advanced digital technology innovation centre, driving early adoption of technologies to make UK businesses more competitive and productive to grow the country's economy.

To learn more:
www.digicatapult.org.uk
www.migarage.ai/ethics-framework/